

(19)



REPUBLIK
ÖSTERREICH
Patentamt

(10) Nummer:

AT 412 476 B

(12)

PATENTSCHRIFT

(21) Anmeldenummer: A 1430/2002
(22) Anmeldetag: 24.09.2002
(42) Beginn der Patentdauer: 15.08.2004
(45) Ausgabetag: 25.03.2005

(51) Int. Cl.⁷: **C12N 15/00**
C12Q 1/68

(56) Entgegenhaltungen:
JINGWEI ET AL. (PNAS 94 (1997), S.
6862-6867
US 6136540
DAIGO ET AL. (DNA RES. 6(4) (1999), S.
227-233
HRABER ET AL. GENOME BIOLOGY 2(9)
(2001): RESEARCH 0037.1-0037.14

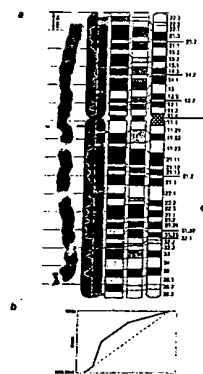
(73) Patentinhaber:
FORSCHUNGSINSTITUT FÜR
KREBSKRANKE KINDER
A-1090 WIEN (AT).
(72) Erfinder:
HAAS OSKAR A. PROF.
WIEN (AT).
SCHMIDT ANDREAS DIPL.ING.
WIEN (AT).

(54) VERFAHREN ZUR HERSTELLUNG EINES VIRTUELLEN CHROMOSOMS

(57) Verfahren zur Herstellung eines virtuellen Chromosoms, das ein entsprechendes natürliches Chromosom repräsentiert, welches Verfahren die folgenden Schritte umfasst:

- Unterteilen von Sequenzdaten des natürlichen Chromosoms in Fraktionen mit einer Länge von mindestens 10.000 bp,
- Bestimmen des CG-Gehalts in jeder Fraktion,
- Berechnen eines Werts zwischen einem Minimalwert und einem Maximalwert für jede Fraktion gemäß dem CG-Gehalt, und
- Herstellen des virtuellen Chromosoms durch Darstellen jeder Fraktion mit dem Wert.

FIG. 3



AT 412 476 B

Die vorliegende Erfindung betrifft ein Verfahren zur Herstellung eines virtuellen Chromosoms, das ein entsprechendes natürliches Chromosom repräsentiert, sowie ein virtuelles Chromosom oder einen Teil davon, das bzw. der durch seinem CG-Gehalt entsprechende Werte dargestellt ist, einen Satz von virtuellen Chromosomen oder Teilen davon und die Verwendung eines Satzes von

5 virtuellen Chromosomen.

Das menschliche Genom ist in einer streng hierarchischen Struktur geordnet. Der Nucleus einer diploiden Zelle enthält etwa $2 \cdot 3,17 \cdot 10^9$ Nucleotidbasen, die in 46 innig verwickelten DNA-Fäden angeordnet sind, welche bei einer Zellteilung in Form von separaten Chromosomen sichtbar werden. Die spezifischen Merkmale dieser Chromosome, wie Zahl, Form, Struktur und Bänderungsmuster, schaffen die Basis für ihre mikroskopische Bestimmung durch herkömmliche zytogenetische Mittel. Diese Methode ist nach wie vor das wichtigste Screening-Werkzeug zur Identifikation von konstitutionellen sowie erworbenen Karyotyp-Anomalien.

Derzeit werden Karyotyp-Anomalien gemäß dem „Internationalen System für Humane Zytogenetische Nomenklatur (ISCN)“, einem System, das auf der diagrammartigen Darstellung von Chromosomen und ihren Bänderungsmustern beruht, beschrieben. Die Bandgrößen und ihre Verteilung in diesen so genannten Ideogrammen leiten sich aus dem Messwert von Trypsin/Giemsa-gefärbten Chromosomenbildern ab, und ihre relative Färbungsintensität wird durch fünf verschiedene Schattierungen symbolisch dargestellt. Da die Anzahl von wahrnehmbaren Bändern auch von der veränderlichen Länge der jeweiligen Chromosome abhängt, stellen Ideogramme mit einer Bandauflösung von 400, 550 und 850 verschiedene Kondensationsstufen dar.

Verglichen mit den hoch ausgeklügelten Computer-Algorithmen und Software-Werkzeugen, die für die Analyse und Evaluierung von molekulargenetischen Daten zur Verfügung stehen, haben sich jene zur Darstellung und Verarbeitung von zytogenetischen Daten innerhalb der letzten 20 Jahre kaum verändert. Einerseits reichte die ISCN-Nomenklatur für die begrenzte räumliche Auflösung der morphologischen Analysen, die resultierende inhärente subjektive Bandzuordnung und die Interpretation der sich daraus ergebenden Anomalien aus. Andererseits verhinderte diese beschreibende Natur bisher eine genauere und objektive Darstellung von zytogenetischen Daten und Daten der fluoreszierenden *in situ* Hybridisierung (FISH) und in der Folge auch ihre nahtlose Integration in bestehende DNA-Datenbanken. Während der letzten zwei Jahrzehnte wurden zahlreiche Meta- und Interphasen-FISH-Technologien wie jene, bei denen heterogene Arten von sequenzspezifischen Sonden, chromosomale Multifarben- und regionsspezifische Färbungs sonden, vergleichende genomische Hybridisierung (CGH) und vergleichende exprimierte Sequenz-Hybridisierung (CESH) verwendet werden, verfügbar und wurden zu Mikroarray-Techniken weiterentwickelt, die rein auf DNA und RNA basieren. Die Beschreibung der sich daraus ergebenden rasch ansammelnden molekulargenetischen Ergebnisse ist gemäß der derzeit verfügbaren zytogenetischen und insbesondere FISH-Nomenklatur mühsam und mit Fehlern behaftet. Die Daten sind schwierig zu verarbeiten, und derzeit ist ihre Integration in eine gemeinsame molekularzytogenetische Datenbank praktisch nicht möglich. Ein Ansatz aus jüngerer Zeit zur Überwindung dieser Hindernisse zumindest zu einem gewissen Grad besteht in der Definition von zytogenetischen Orientierungspunkten mit Hilfe von homogen beabstandeten FISH-Ankersonden entlang den Chromosomen und den entsprechenden Ideogrammen.

Es wird schon lange als gegeben anerkannt, dass das chromosomale Bänderungsmuster abwechselnd CG-reiche und CG-arme Sequenzabschnitte widerspiegelt. Nichtsdestotrotz herrschte die Ansicht vor, dass die Korrelation zwischen Chromosomenbändern und dem CG-Gehalt insgesamt nur als ziemlich schwache Approximation betrachtet werden kann. In diesem Zusammenhang schien es höchst unwahrscheinlich, dass das Bänderungsphänomen einfach das Ergebnis von langbereichigen Veränderungen in der linearen Basenpaarzusammensetzung allein sein sollte. Es wurde vielmehr angenommen, dass das Bänderungsmuster signifikant von strukturellen Faktoren wie Faltung, Proteindeckung, DNA-Packung und -Kondensation sowie der Erreichbarkeit der DNA durch Farben mitbestimmt und modifiziert wird.

Schließlich bekräftigte ein direkter Berechnungsvergleich zwischen dem sequenzspezifischen CG-Gehalt und dem speziellen Färbungsmuster, das jüngst möglich wurde, diesen Zusammenhang weiter (Niimura und Gojobori („*In silico* chromosome staining: Reconstruction of Giemsa bands from the whole human genome sequence”, PNAS, Band 99, Nr. 2 (797-802)). Ihre „*in-silico*-Chromosomenfärbung“ wurde mit Hilfe eines Verfahrens mit zwei Fenstern, einem lokalen Fenster

von 2,5 mb und einem regionalen Fenster von 9,3 mb, erreicht, wobei die Beziehung zwischen dem CG-Gehalt im lokalen Fenster in Bezug auf den GC-Gehalt im regionalen Fenster berechnet wurde. Nach Niimura und Gojobori würde dieses Zweifenster-Verfahren besser eine in-silico-Färbung bewirken als die Giemsa-Bänderungsmuster nur durch den Unterschied in der Basenzusammensetzung zu erklären. Weiters wird angenommen, dass sich der Leistungsgrad verbessert würde, wenn der Unterschied eines Verdichtungsverhältnisses zwischen G- und R-Bändern berücksichtigt würde, wobei G-Bänder stärker kondensiert werden als R-Bänder. Durch eine Berechnung von 10-kb-Fragmenten der menschlichen DNA-Sequenz und verschiedene Möglichkeiten statistischer Analysen fanden sie heraus, dass bei einem Bandniveau von 850 der CG-Gehalt und die Giemsa-Bänder entlang eines Chromosoms am besten bei einem lokalen Fenster von 2,5 Mb und einem regionalen Fenster von 9,3 Mb korrelierten. Die Größe dieser Fenster wurde so gewählt, dass die Übereinstimmung zwischen in-silico- und Giemsa-Bändern optimiert wurde, doch stellten die Autoren auch fest, dass ihr Ansatz für feine Bänder, die kleiner als das lokale Fenster sind, möglicherweise nicht geeignet wären. Sie gingen jedoch davon aus, dass die Korrelation zwischen Giemsa- und in-silico-Bändern durch eine Integration der Genom-weiten FISH-Kartierungsdaten weiter verbessert werden könnte. In dieser Veröffentlichung wurden jedoch keine Chromosome kondensiert, sondern Ideogramme berechnet und verglichen. Nach Niimura und Gojobori zeigen ihre Ergebnisse, dass Giemsa-Bänderungsmuster nur durch den Unterschied in der Basenzusammensetzung nicht erklärt werden können. Die Beziehung zwischen der Nukleotidsequenz und zygoten Bändern würde somit nach wie vor illusorisch bleiben.

Frühere Versuche, die zytogenetische Karte mit der Sequenz des menschlichen Genoms zu verknüpfen, konzentrierten sich auf einen Top-Down-Ansatz entweder durch die Abgrenzung von Bändergrenzen oder durch das Setzen von bandunabhängigen zytogenetischen Orientierungspunkten mit spezifischen FISH-Sonden. Beispielsweise platzierte das „BAC Resource Consortium“ 7.600 solcher zytogenetisch definierter Orientierungspunkte auf dem Sequenzentwurf des menschlichen Genoms. Zwar sollten diese Marker unter anderem „auch eine strikte Beurteilung von Sequenzunterschieden zwischen den dunklen und hellen Bänder von Chromosomen zulassen“, doch wird ihre Lage nach wie vor auf Ideogrammen veranschaulicht, die nicht mit der DNA-Sequenz verknüpft sind. Desgleichen stellt das „Krebschromosomen-Aberrationsprojekt“ (Cancer Chromosome Aberration Project - CCAP) ein zytogenetisches Koordinatensystem mit willkürlich definierten Intervallen auf Basis eines Ideogramms auf, um die Lage von Sequenz-verankerten FISH-Klonen aufzuzeigen.

In *Jingwei et al.* (PNAS 94 (1997), S. 6862-6867) wird allgemein der GC-Gehalt von Chromosomenteilen untersucht.

In der *US 6 136 540 A* wird ein Computerprogramm zum Auffinden genetischer Abnormitäten beschrieben, mit dem die subjektive Analyse von selektiv gefärbten Chromosomen vermieden werden kann. Gemäß diesem Dokument werden die chromosomalen Abnormitäten durch spezifische Hybridisierungssonden (die mit Fluorophoren markiert sind) bestimmt und analysiert, wobei chromosomale Additionen, Deletionen, Amplifikationen, Translokationen und Umbildungen erkannt werden können. Obgleich gemäß diesem Dokument die Nachteile experimenteller Chromosomenfärbungen vermieden werden sollen bzw. deren Probleme bei der Reproduzierbarkeit, wird auch hierbei experimentell aufwendig, nämlich mit Fluorophor-markierten Hybridisierungssonden, gearbeitet, wodurch ebenfalls wiederum Reproduktionsschwierigkeiten und die experimentellen Nachweisen inhärenten Ungenauigkeiten vorhanden sind.

Gemäß *Daigo et al.* (DNA Res. 6(4) (1999), S. 227-233) wird der GC-Gehalt auf bestimmten Abschnitten auf Chromosom 9 und 3 bestimmt und zwar nach konventionellen Methoden.

Der Artikel von *Hrabec et al.* (Genome Biology 2(9) (2001), research 0037.1-0037.14) betrifft eine Analysemöglichkeit zur Untersuchung von interspezifischen Wechselwirkungen in Sequenzen, die während der Wechselwirkung zwischen zwei Symbionten exprimiert werden, z.B. hinsichtlich ihres GC-Gehaltes. Dabei werden aber weder komplette größere Bereiche des Genoms miteinander verglichen, noch konkret als Chromosomen dargestellt.

Ein Ziel der vorliegenden Erfindung ist daher die Schaffung eines Verfahrens zur Herstellung eines virtuellen Chromosoms, das ein genaues Bänderungsmuster auf skalenuabhängige und stark bereichsspezifische Weise mit hoher Auflösung ermöglicht. Weiters sollten diese hergestellten virtuellen Chromosome nicht nur morphologische Informationen wie die herkömmlichen Ideogram-

me oder ISCN-Bänder, sondern auch die entsprechenden genetischen Informationen, z.B. die Sequenzdaten, aufweisen. Ein solches virtuelles Chromosom, das die kompletten Sequenzdaten aufweist, mit den herkömmlichen Darstellungen von Chromosomen vergleichbar ist und eine hohe Auflösung hat, wurde bisher noch nicht erzeugt. Daher besteht ein weiteres Ziel der vorliegenden

5 Erfindung in der Schaffung eines Chromosomensatzes, der als Schnittstelle zwischen herkömmlichen Darstellungen von Chromosomen und genetischen Informationen und Sequenzdaten verwendet werden kann, um von einer DNA-Sequenz abgeleitete Daten und natürliche chromosomale Bänderungsmuster direkt zu vergleichen.

Das Ziel der vorliegenden Anmeldung wird durch ein erfindungsgemäßes Verfahren wie oben definiert erreicht, welches dadurch gekennzeichnet ist, dass es die folgenden Schritte umfasst:

- Unterteilen von Sequenzdaten des natürlichen Chromosoms in Fraktionen mit einer Länge von mindestens 10.000 bp,
- Bestimmen des CG-Gehalts in jeder Fraktion,
- Berechnen eines Werts zwischen einem Minimalwert und einem Maximalwert für jede Fraktion gemäß dem CG-Gehalt, und
- Herstellen des virtuellen Chromosoms durch Darstellen jeder Fraktion mit dem Wert.

Es stellte sich heraus, dass mit diesem Verfahren hergestellte Chromosome eine hervorragende Korrelation zwischen dem eigenen Bänderungsmuster und dem ihrer entsprechenden natürlichen Gegenstücke lieferten. Diese überraschende Übereinstimmung zeigt nicht nur, dass das chromosomale Bänderungsmuster zu einem Großteil direkt durch die zugrunde liegende DNA-Sequenz bestimmt wird, sondern kann auch eine einzigartige Basis für die gemeinsame Verarbeitung von morphologischen und molekulargenetischen Daten innerhalb eines einzigen, auf der DNA-Sequenz basierenden Gerüsts bieten. Im Gegensatz zu aktuellen Publikationen, die ausdrücklich feststellen, dass Giemsa-Bänderungsmuster durch die unterschiedliche Basenzusammensetzung allein nicht erklärt werden können, wird mit dem vorliegenden Verfahren gezeigt, dass eine Darstellung von Chromosomen auf Sequenzbasis gemäß dem CG-Gehalt sehr wohl möglich ist und zu virtuellen Chromosomen mit hoher Auflösung führt.

Im Rahmen der vorliegenden Anmeldung bezieht sich der Ausdruck „Verfahren zur Herstellung eines virtuellen Chromosoms, das ein entsprechendes natürliches Chromosom repräsentiert“ nicht nur auf vollständige Chromosome, sondern auch auf Teile davon, beispielsweise separate Chromosomenarme oder -enden. Für die Sequenzdaten des natürlichen Chromosoms können beispielsweise irgendwelche elektronisch verfügbaren Daten herangezogen werden, im Fall von menschlichen Chromosomen können das zum Beispiel Sequenzen aus dem Arbeitsentwurf des Humangenomprojekts [Human Genome Project Working Draft (<http://genome.ucsc.edu/>)] sein. Im Rahmen der vorliegenden Anmeldung bezieht sich der Ausdruck „Chromosom“ auf irgendwelche Chromosome irgendeines Organismus. Der Organismus ist zum Beispiel der Mensch, es kann jedoch auch irgendein Lebewesen, insbesondere ein Säuger, den Organismus darstellen, aus dem das Chromosom stammt. Virtuelle Chromosome von Säugern und Menschen sind besonders bevorzugt, da sie für Evolutionsstudien verwendet werden können.

Der Schritt des Unterteilens der Sequenzdaten in Fraktionen und des Bestimmens des CG-Gehalts in jeder Fraktion wird vorzugsweise elektronisch ausgeführt.

Der Ausdruck „Wert zwischen einem Minimalwert und einem Maximalwert“ bezieht sich auf irgendeinen Parameter, der geeignet ist, einen spezifischen CG-Gehalt zu definieren und vorzugsweise virtuell darzustellen. Das kann beispielsweise ein Prozentwert zwischen 0% und 100% oder ein Wert zwischen 0 und 1 sein, der zum Beispiel in einem zwei- oder dreidimensionalen Bild sichtbar gemacht wird. Die Werte können weiters durch Lichtwerte oder Farbwerte dargestellt werden. Jeder Wert, der sichtbar gemacht werden kann, ist zur Darstellung eines bestimmten CG-Gehalts geeignet.

Der Schritt des Berechnens des Werts für jede Fraktion gemäß ihrem CG-Gehalt kann mit jeder geeigneten Tabelle oder Formel, jedem geeigneten Algorithmus oder Programm durchgeführt werden, wobei zum Beispiel ein Minimalwert einer Minimalmenge an CG in einer Fraktion zugeordnet wird und ein Maximalwert einer Maximalmenge an CG in einer Fraktion zugeordnet wird. Die Werte dazwischen werden dann als lineare Funktion zwischen den beiden Extremwerten dem wechselnden CG-Gehalt in den Fraktionen zugeordnet.

Vorzugsweise ist der Wert ein Lichtwert. Der Vorteil eines Lichtwerts besteht darin, dass die

Sichtbarmachung sehr rasch und einfach interpretiert werden kann und weiters mit konventionell erzeugten, beispielsweise mikroskopisch entnommenen oder ISCN-Chromosomen verglichen werden kann.

Auf weiters bevorzugte Weise werden der Maximalwert weiß, der Minimalwert schwarz und Werte dazwischen in Grauschattierungen dargestellt. Diese Darstellung eines virtuellen Chromosoms ist direkt vergleichbar mit den herkömmlich gescannten Chromosomen, die Auflösung ist jedoch sehr hoch und das virtuelle Chromosom umfasst weiters die Sequenzdateninformationen, die in den Chromosomendarstellungen des Standes der Technik fehlen.

Gemäß einer bevorzugten Ausführungsform wird das natürliche Chromosom in Fraktionen einer Länge bis 1.000.000 bp, vorzugsweise einer Länge von 50.000 bis 500.000 bp, noch bevorzugter einer Länge von 100.000 bis 300.000 bp unterteilt. Diese Fraktionen sind ausreichend klein, um hohe Auflösungen und eine größtmögliche Fülle an Informationen zu ermöglichen. Vorzugsweise entspricht eine Fraktion der geschätzten Durchschnittsgröße einer DNA-Schleife sowie jener eines Isochors. Eine optimale Fraktionslänge beträgt beispielsweise 200.000 Basenpaare.

Vorteilhaft wird die Fraktion mit einem CG-Gehalt von 30 bis 35 %, vorzugsweise 33 %, einem Minimalwert und die Fraktion mit einem CG-Gehalt von 60 bis 65 %, vorzugsweise 62 %, einem Maximalwert zugeordnet. Es wurde gefunden, dass eine Veränderung zwischen diesen Prozentangaben Chromosome mit grauen Bandwerten erzeugt, die den herkömmlich dargestellten Chromosomen entsprechen. Daher kann diese Sichtbarmachung des virtuellen Chromosoms direkt mit herkömmlichen Chromosomendarstellungen verglichen werden kann, wie sie beispielsweise durch ein Mikroskop wahrgenommen werden.

Vorzugsweise wird Fraktionen mit unbekannter Sequenz der Wert gemäß ihrer morphologischen Erscheinung zugeordnet. Auch wenn die Menge an Fraktionen, denen eine Sequenz fehlt, insbesondere bei menschlichen Chromosomen aufgrund des nahezu vollständigen menschlichen Genoms sehr gering geworden ist und im Allgemeinen rasch abnehmen wird, können die wenigen Fraktionen mit fehlender Sequenz durch Daten ergänzt werden, die von der morphologischen Erscheinung abgeleitet, beispielsweise einem Ideogramm entnommen worden sind, um ein vollständiges Chromosom zu liefern.

Auf weiters bevorzugte Weise wird nach der Herstellung des virtuellen Chromosoms ein Filter zur Glättung der Erscheinung, vorzugsweise ein Gaußsches Faltungsfilter, angewendet. Dabei wurden auch die resultierenden Schattierungen verwendet, um die wenigen letzten Pixels an den Chromosomengrenzen nach und nach zu füllen.

Gemäß noch einem bevorzugten Verfahren wird zur Herstellung des virtuellen Chromosoms ein Skalenkorrekturfilter angewendet. Das kann ein Normalisierungs- und nichtlineares Grauskalenkorrekturfilter vom Gamma-Typ sein. Dadurch kommt es zu einer Kontrastverstärkung und das Bild von Chromosomen, wie sie durch das Mikroskop wahrgenommen werden, wird optimal nachgeahmt.

Ein weiterer Aspekt der vorliegenden Anmeldung betrifft ein virtuelles Chromosom oder einen Teil davon, das bzw. der durch Werte gemäß seinem CG-Gehalt dargestellt wird, welcher bzw. welches dadurch gekennzeichnet ist, dass die Herstellung gemäß dem oben definierten erfindungsgemäßen Verfahren erfolgt. Gemäß der vorliegenden Erfindung wird ein virtuelles Chromosom geschaffen, das nicht nur morphologische Informationen umfasst und beispielsweise mit herkömmlichen Ideogrammen verglichen werden kann, sondern welches auch Sequenzdaten enthält. Dieses auf Sequenzbasis visualisierte Chromosom gemäß der vorliegenden Erfindung zeigt eine hervorragende Korrelation des Bänderungsmusters von virtuellen Chromosomen und jenem der entsprechenden natürlichen Gegenstücke. Hinsichtlich dieses Aspekts der vorliegenden Erfindung gelten dieselben Definitionen und bevorzugten Ausführungsformen wie oben.

Vorzugsweise ist der Wert ein Lichtwert, noch bevorzugter sind der Maximalwert weiß, der Minimalwert schwarz und Werte dazwischen in Grauschattierungen dargestellt. Wie oben ausgeführt, ermöglicht dies eine Darstellung des Chromosoms, wie es durch das Mikroskop gesehen wird, und ist daher optimal für einen Vergleich mit herkömmlichen Chromosomendarstellungen, wie beispielsweise Ideogrammen oder mikroskopisch wahrgenommenen Chromosomen, geeignet.

Gemäß einem weiteren Aspekt der vorliegenden Anmeldung wird ein Satz von virtuellen Chromosomen oder Teilen davon zur Verfügung gestellt, der dadurch gekennzeichnet ist, dass er zwei oder mehr erfindungsgemäße Chromosome oder Teile davon wie oben definiert umfasst.

Vorzugsweise umfasst der Satz eine Höchstzahl von Chromosomen, wobei dieser Satz ständig durch weitere, neu gefundene oder neu identifizierte Chromosome ergänzt werden kann.

Vorzugsweise umfasst der Satz Chromosome oder Teile davon, die für einen oder mehrere Organismen spezifisch sind. Der Vorteil eines Satzes, der für einen Organismus spezifisch ist, liegt darin, dass dieser Satz für einen Vergleich von irgendwelchen neu identifizierten Modifikationen oder Umlagerungen von Chromosomen dieses Organismus nützlich ist. Es ist jedoch selbstverständlich möglich, einen Satz mit Chromosomen verschiedener, vorzugsweise definierter Organismen zur Verfügung zu stellen.

Auf weiters bevorzugte Weise umfasst der Satz 24 menschliche Chromosome oder Teile davon. Dieser Satz ist ein Standardsatz für normale menschliche Chromosome und kann für einen Vergleich von Chromosomen eines Patienten mit normalen Chromosomen zum Nachweis irgendwelcher Modifikationen oder Umlagerungen verwendet werden.

Auf weiters bevorzugte Weise umfasst der Satz weiters zusätzliche modifizierte Chromosome oder Teile davon, vorzugsweise Chromosome mit Translokationen. Das ist von besonderem Vorteil für modifizierte Chromosome, die mit einer spezifischen Krankheit, beispielsweise einem spezifischen Tumor, verwandt sind. Indem eine Klassifikation von solchen modifizierten virtuellen Chromosomen, vorzugsweise jeder Modifikation mit einem Hinweis auf eine spezifische Krankheit, vorgesehen wird, ist es ganz leicht möglich, eine Krankheit oder die Gefahr des Ausbrechens einer Krankheit einem Satz von aus einem Patienten isolierten Chromosomen zuzuordnen, indem der Satz von virtuellen Chromosomen mit den Chromosomen des Patienten verglichen wird. Aufgrund des ständigen Nachweises neuer Modifikationen in Chromosomen kann der Satz rasch und permanent mit den neuesten medizinischen Informationen vervollständigt werden.

Im Rahmen der vorliegenden Anmeldung bezieht sich der Ausdruck „Chromosomenmodifikation“ auf irgendeine Sequenzmodifikation, z.B. irgendeine Mutation oder Translokation eines Chromosomenfragments.

Ein weiterer Aspekt der vorliegenden Erfindung ist die Verwendung des erfindungsgemäßen Satzes von virtuellen Chromosomen wie oben genannt zur Katalogisierung von Chromosomenmodifikationen. Wie oben ausgeführt, ist der erfindungsgemäße Satz besonders zweckmäßig zur Lieferung von elektronischen Informationen über Chromosomenmodifikationen und ihren Konnex zu irgendeiner Krankheit oder Krankheitsgefahr. Im Rahmen der vorliegenden Anmeldung bezieht sich der Ausdruck "Chromosomenmodifikation" auf irgendeine Modifikation im Chromosom. Es kann sich dabei um eine Sequenzmutation oder die vollständige Translokation eines Chromosomenfragments handeln. Aufgrund der hohen Auflösung kann jede Chromosomenmodifikation detektiert und katalogisiert werden.

Der erfindungsgemäße Satz gestattet die Beschreibung von Chromosomenanomalien mit einer bisher unbekannten molekularen Präzision, wobei andererseits nach wie vor die Möglichkeit besteht, verschwommenere Großereignisse auf reiner Chromosomenebene zu interpretieren, wie es auch mittels konventioneller zytogenetischer Analyse sowie mittels vergleichender Genomhybridisierung und vergleichender exprimierter Sequenzhybridisierung auf Basis von chromosomaler Multifarbenfärbung FISH möglich ist.

Ein weiterer Aspekt der vorliegenden Anmeldung betrifft die Verwendung eines erfindungsgemäßen Satzes von virtuellen Chromosomen wie oben definiert zur virtuellen Kartierung der chromosomalen Position einer Sequenz. Da der erfindungsgemäße Satz von virtuellen Chromosomen von der vollständigen menschlichen DNA-Sequenz abgeleitet ist bzw. diese darstellt, ist es möglich, die chromosomale Position irgendeiner gegebenen, bekannten oder unbekannten Sequenz oder Sequenzgruppe, die in einer Datenbank zur Herstellung der virtuellen Chromosome enthalten ist, zu kartieren und anzuzeigen.

Ein weiterer Aspekt der vorliegenden Anmeldung bezieht sich auf die Verwendung eines erfindungsgemäßen Satzes von virtuellen Chromosomen wie oben definiert als Schnittstelle zwischen morphologischen und molekulargenetischen Daten. Vorzugsweise stammen die morphologischen Daten von Informationen basierend auf dem Internationalen System für Humane Zytogenetische Nomenklatur (ISCN).

Dabei kann eine grafische Schnittstelle über jede molekulargenetischen Datenbank überlagert werden. So gewährleistet ein virtuelles System auf Chromosomenbasis, dass zuvor erfasste Daten zugänglich und analysierbar bleiben. Beispielsweise ist eine Kombination von grafischen Schnitt-

stellen-Werkzeugen auf Basis von ISCN-Nomenklatur und virtuellen Chromosomen, die bestehende zytogenetische Datenbanken überlagern kann wie oben ausgeführt, äußerst nützlich. Eine derartige Schnittstelle ermöglicht die Transformation von ISCN-Informationen in das entsprechende Karyotyp-Bild. Umgekehrt kann ein Karyotyp-Bild, das mit einem solchen virtuellen Chromosomen-Werkzeug generiert wird, in ein ISCN-Band translatiert werden. Eine derartige grafische Schnittstelle ist extrem wertvoll für eine visuelle Kreuzprüfung der ISCN-Beschreibung durch einen Vergleich des Karyotyp-Bilds mit dem virtuellen Chromosomenbild. Als wertvolle Nebenerscheinung verbessert ein solcher Ansatz auch signifikant die Qualität von zytogenetischen Daten. Darüber hinaus erleichtert er auch den nahtlosen Austausch und Verkehr von zytogenetischen Daten in standardisierter Form in einem Labor bzw. zwischen Labors nicht nur mit einer Fernzentrale, sondern auch mit FISH- und molekulargenetische Datenbanken. Beispielsweise können dann zytogenetische Daten, die zur Veröffentlichung vorbereitet werden, leicht überprüft und bequem an eine zentrale Datenbank übermittelt werden.

Die Überlagerung von virtuellen Chromosomen als grafische Schnittstelle über molekulargenetische Datenbanken unterstützt die Visualisierung jeder Art von FISH-, DNA- und RNA-abgeleiteten Datensätzen sowie Gen-Expressionsprofilen auf standardisierte „chromosomale“ Weise. Die Vorteile einer solchen chromosomal Darstellung liegen darin, dass diese unabhängig von der Sondenverteilung auf den verschiedenen Arrays ist, und auch darin, dass ihre „natürliche“ Erscheinung das Verständnis und einen Vergleich durch visuelle Prüfung erleichtert. Darüber hinaus kommt einer derartigen Darstellung von Gen-Expressionsprofilen immer mehr Bedeutung zu, weil es zunehmend Beweismittel dafür gibt, dass auch funktionell nichtverwandte Gene in Transkriptionsterritorien in *Drosophila* sowie im menschlichen Genom exprimiert werden. Das resultierende Verteilungsmuster gleicht jenen, die von CGH- und CESH-Analysen stammen, bei denen unterschiedlich markierte DNA oder cDNA aus einem Gewebe von Interesse und eine Kontrollprobe gleichzeitig direkt auf Chromosome hybridisiert werden. In der Folge können solche Datensätze mit anderen Karyotyp-Mustern, beispielsweise den zugehörigen Karyotyp-Anomalien, direkt korreliert und kreuzanalysiert werden.

Da die Position von FISH- oder anderen DNA-Sonden (von Oligonucleotiden bis regionsspezifischen Färbungssonden) bequem auf virtuellen Chromosomen angezeigt werden und mit den tatsächlich erhaltenen Hybridisierungsmustern kreuzgeprüft werden kann, sind solche grafischen Schnittstellen auch von potentiell Interesse für Ressourcen-Zentren. Darüber hinaus kann die Position von zytogenetischen Orientierungspunkten in Form von gleichmäßig verteilten FISH-Sonden für Zwecke des Kartierens von Defektstellen direkt in derartige Chromosome integriert werden. Schließlich können sogar submikroskopische Ereignisse, die andernfalls mit konventionellen zytogenetischen Mitteln nicht nachweisbar sind, wie Mikrodeletionen und Interphasen-FISH-Daten, kartiert und in eine solche universelle Plattform inkludiert werden.

Vorzugsweise dient der Chromosomensatz als Referenz zur Klassifizierung eines Phänotyps zu einer Sequenzanordnung. Der Ausdruck „Sequenzanordnung“ bezieht sich auf irgendeine Modifikation, z.B. Sequenzmutationen oder Translokationen von Chromosomenfragmenten. Der Phänotyp kann sich auf normale oder abnormale Phänotypen beziehen, z.B. verschiedenen Krankheiten wie Tumore. Insbesondere bei einer Klassifizierung der Chromosome nach Modifikationen und resultierenden Phänotypen kann jedes aus einem Patienten isolierte und mit herkömmlichen mikroskopischen Methoden analysierte Chromosom mit dem erfindungsgemäßen Chromosomensatz verglichen werden. Ähnlichkeiten zwischen den Modifikationen der Chromosome würden auch ähnliche Phänotypen oder zumindest die Wahrscheinlichkeit oder das Risiko der Entstehung eines ähnlichen Phänotyps implizieren.

Vorteilhaft dient der Chromosomensatz als Werkzeug zur Durchführung von Struktur- bzw. Funktionsanalysen einer Sequenzanordnung. Wie oben ausgeführt, kann die Analyse durch Genkartierung oder virtuelle Hybridisierung aufgrund der im virtuellen Chromosom enthaltenen Sequenzdaten durchgeführt werden.

Auf weiters bevorzugte Weise dient der Chromosomensatz als Werkzeug zur Bestimmung des Einflusses eines bestimmten Faktors auf eine Sequenzanordnung. Beispielsweise kann ein externer Faktor wie eine chemische Substanz, Energie mit verschiedenen Wellenlängen oder auch der Einfluss von Mikroorganismen auf zytogenetischer Basis analysiert und in den erfindungsgemäßen Chromosomensatz transferiert oder mit diesem verglichen werden, wodurch Implikationen oder

resultierende Phänotypen abgeleitet oder vorhergesehen werden können.

Die vorliegende Erfindung wird genauer anhand der folgenden Beispiele und Figuren beschrieben, auf die sie jedoch nicht beschränkt ist, worin:

Fig. 1 Bilder einer von menschlichen Chromosomen abgeleiteten DNA-Sequenz im Vergleich zu Trypsin/Giemsa-gebänderten Chromosomenbildern zeigt;

Fig. 2 ein Modell von ungleichmäßiger Kondensation von G-Bänder in Abhängigkeit von ihrem CG-Gehalt darstellt;

Fig. 3 virtuelle Chromosome im Vergleich zu den zytogenetischen und molekulargenetischen Karten zeigt;

Fig. 4 eine virtuelle in-situ-Hybridisierung zeigt;

Fig. 5 die Konstruktion von virtuellen Chromosomenabnormalitäten zeigt; und

Fig. 6 eine grafische Schnittstelle zwischen zytogenen und molekulargenetischen Datensätzen darstellt.

Beispiele

Beispiel 1:

Herstellung von virtuellen Chromosomen auf der Basis von genetischen Daten

Zur Herstellung von virtuellen Chromosomen wurden die Sequenzdaten der August- und Dezember-2001- sowie April- und Juni-2002- Ausgaben des Humangenom-Arbeitsentwurfs (human genome working draft) heruntergeladen. Zur Analyse und Verarbeitung der Daten und der daraus resultierenden Bilder wurden Perl, Mathematica (Wolfram Scientific) und Photoshop (Adobe) verwendet. Unter Verwendung der Script-Sprache Perl wurden zuerst die Sequenzdaten jedes einzelnen Chromosoms in 200.000 Basen lange Fraktionen unterteilt. Diese Größe entspricht der geschätzten Durchschnittsgröße einer DNA-Schleife und auch annähernd jener eines Isochors. Dann wurde der CG-Gehalt aller Streifen prozentuell ermittelt, deren Sequenz zu mindestens 70 % vollständig war. Das traf praktisch in allen Fällen beginnend mit der Dezember-2001-Ausgabe zu. Lücken aufgrund von nicht sequenzierten Nucleotiden (N's) wurden nicht in Betracht gezogen. Der CG-Gehalt der einzelnen Streifen reichte von 33 % bis 62 % bei einem Mittelwert von 41 %. Die Tabellen mit diesen Daten wurden zusammen mit den Informationen über die Segment- und Bandkoordinaten des jeweiligen Chromosoms in temporären Files gespeichert und mit Mathematica weiter berechnet und assembliert wie nachstehend gezeigt:

```
StaticNormalize[L_List]:= Block [{Lx = 0,625, Li = 0,33},  $\left( \frac{\#1 - Li}{Lx - Li} \right) / @L ]$ 
```

```
StaticNormalize[L_Real]:= Block [{Lx = 0,625, Li = 0,33},  $\frac{L - Li}{Lx - Li} ]$ 
```

Entsprechend dem einzelnen CG-Gehalt wurde jedem Streifen ein Grauwert auf lineare normalisierte Weise zugeordnet, d.h. Streifen mit einem CG-Gehalt von 33 % wurden schwarz und jene mit einem CG-Gehalt von 62 % weiß. Die Übertragungsfunktion in Mathematica von Prozent CG auf den Grauwert ist die Summe aus (statistischer) Normalisierung und Kontrastverstärkung:

```
BandsAvg = ListConvolve[FoldMask, StaticNormalize[ChrShades],  
  {CenterElement, -CenterElement}, .41] / (Plus @@ FoldMask)
```

Die abgeleiteten Balken wurden dann innerhalb der jeweiligen Chromosomengrenzen, wie durch ihre Länge und zentromere Position definiert, integriert. Die zentromeren, heterochromatischen und Satelliten-Regionen, für die noch keine geeigneten Sequenzinformationen erhältlich sind, wurden entsprechend ihrer morphologischen Erscheinung künstlich ergänzt. Zur Glättung der Erscheinung der virtuellen Chromosome wurde ein Gaußsches Faltungsfilter (N(0,1), 22 Streifen in der Länge) angewendet:

```
NiceUnitBand[Pos_, Width_, Stain_, BitFields_, Y_, H_, ChrNames_] := Block[{CL},  
  CL = Select[{ChrStartPos[ChrNames], ChrSatelitePos[ChrNames]} ~ Join ~  
  ChrCentroPosList[ChrNames] ~ Join ~ {ChrEndPos [ChrNames] }, NumberQ];
```

```

{
  Raster[Table[If[False, {Max[0, Min[1, Schnitt[Stain] [x]]}],
    {Max[0, Min[1, Schnitt[ColorCorrection[Stain] [x]]}], (x, 0, 1,  $\frac{1}{30}$ )},
    {{Pos - Width/2, Y + H * (1 - Boundary[CL, Pos])},
    {Pos + Width/2, Y + H - H * (1 - Boundary[CL, Pos])}},
    ColorFunction → GrayLevel
  ],
  MapIndexed [
    Rectangle[{Pos - Width/2, - (#2[[1]] * 10 + 1)}, {Pos + Width/2, - (#2[[1]] * 10 + 9)},
    Graphics[{ , Background → If [#1 >= 0, Hue[#1], GrayLevel[1] ] ]
  ] &, BitFields ] ]

```

Weiters wurden die resultierenden Schattierungen dazu verwendet, nach und nach die wenigen letzten Pixels an den Chromosomengrenzen zu füllen. Zur Kontrastverstärkung und Nachahmung der Chromosomenbilder, wie sie durch ein Mikroskop wahrgenommen werden, wurde ein Normalisierungs- und nichtlineares Grauskalen-Korrekturfilter vom Gamma-Typ verwendet:

```

ColorCorrection[c_]:= (* 1+(c-1)^ 3*) ·
Interpolation[{{0,0}, { .1, .1), { .2, .5), { .5, .8), {1, 1}},
InterpolationOrder → 1] [c]

```

Um die Daten in Chromosomenform zu bringen:

```

Boundary[CL_, p_]:= Block[{r = 5000000, d},
d = Min[Abs[CL-p]];

```

```

If[d ≤ r,  $\frac{1}{3} \sqrt{1 - \left(1 - \frac{d}{r}\right)^2} + N[2/3, 1]$  ]

```

Als letzter Schritt wurden die jeweiligen Bilder dann in Photoshop importiert, assembliert und entsprechend arrangiert.

Beispiel 2:

Bilder von DNA-Sequenz-abgeleiteten menschlichen Chromosomen

In Fig. 1 sind Trypsin/Giemsa-gebänderte Chromosomenbilder dargestellt, wobei für jedes Chromosom (a) ein ISCN-Referenzbild mit 850 Bänderstufen zeigt, (b) ihr abgeleitetes gerade gerichtetes Grauschattierungsmuster zeigt, (c) ein Vergleich mit ihren berechneten virtuellen Gegenständen der August-2001- und (d) Dezember-2001-, (e) April-2002- und (f) Juni-2002-Ausgabe ist (siehe <http://genome.ucsc.edu/>).

Trotz der hervorragenden allgemeinen Gesamtkonkordanz zwischen den gematchten Sätzen von Chromosomen-Homologen wurden einige örtliche Veränderungen und Unterschiede offenkundig, insbesondere zwischen virtuellen Chromosomen, die von unterschiedlichen Sequenzausgaben stammen. Die ursprünglich hervorragende Übereinstimmung zwischen dem Grauskalen-Bänderungsmuster der natürlichen Chromosome und der virtuellen Chromosome der Ausgabe vom August 2001 verschlechterte sich bei Verwendung der Ausgabe vom Dezember 2001 für Vergleichszwecke. Diese verblüffende Feststellung lässt sich dadurch erklären, dass die spätere Zusammenstellung bei CNBI und nicht bei USCS produziert wurde. Bei einem Vergleich mit der UCSC-Zusammenstellung zeigt die NCBI-Zusammenstellung eine etwas bessere lokale Ordnung und Orientierung, jedoch eine etwas schlechtere Verfolgung der Chromosomenniveauearten. Solche Verschiebungen von Sequenzabschnitten können somit das örtliche Bänderungsmuster merklich verändern, was besonders deutlich wird bei einem Vergleich der langen Arme der virtuellen Chromosome 1 und 11 von verschiedenen Ausgaben. Die Lage der Zentromere der Chromosome 5 (Ausgabe Dezember 2001), 7 und 12 (beide Ausgabe 2002) verschob sich zu ungeraden Positionen. Bemerkenswert ist jedoch, dass sich die Vorlagerungen und ständigen Korrekturen in

der Sequenzzusammenstellung ebenso signifikant verbesserten wie die Konkordanz zwischen den natürlichen und virtuellen Bänderungsmustern. Ein derartiger Vergleich von virtuellen Chromosomen, die von verschiedenen Sequenzausgaben stammen, kann somit auch eine unabhängige Validierung der Sequenzkarte ermöglichen.

5

Beispiel 3:

Modellierung von ungleichmäßiger Kondensation von G-Bänder in Abhängigkeit von ihrem CG-Gehalt

Virtuelle Chromosome gestatten Struktur- und Funktionsanalysen des Genoms und schaffen Möglichkeiten zur Untersuchung des Einflusses verschiedener Faktoren auf das groß angelegte chromosomale Bänderungsmuster. Das hierin gezeigte Beispiel bezieht sich auf potentielle Wirkungen der ungleichen Kontraktion von hellen und dunklen Bändern während der Chromosomenkondensation. Die Analyse beruht auf der Vorstellung, dass dunkle Giemsa-Bänder bis zu elfmal mehr DNA enthalten können als die hellen Bänder und dass das DNA-Verdichtungsverhältnis in der Größenordnung der Kubikwurzel der jeweiligen DNA-Länge liegt (siehe Fig. 2).

Zuerst wurden die Bilder des kürzesten (ISCNS, Bandstufe 500) und längsten (ISCNL, Bandstufe 850) ISCN-Chromosoms 7 in ein Grauskalenmuster durch Messen der Grauwerte entlang des blauen Pfads (a) transformiert. Nachdem diese zwei Chromosomenbilder auf dieselbe Länge gebracht worden waren, wurde ihr Bänderungsmuster mit jenem von virtuellen Homologen verglichen, die auf verschiedene Arten modifiziert wurden. In Abhängigkeit vom jeweiligen CG-Gehalt, und wie in der Grafik erläutert, (s. Fig. 2b) wurde die Länge der hellen und dunklen Bänder der virtuellen Chromosome linear gedehnt oder kondensiert, indem die Faktoren 0,3, 0,5 bzw. 0,8 angewendet wurden, die annähernd einem 2,2-, 3,4- bzw. 5,8-fachen Unterschied in ihrer DNA-Länge entsprechen (s. Fig. 2c). Ein Vergleich der resultierenden Bänderungsmuster bestärkt frühere experimentelle Nachweise, dass die Chromosomenkondensation aller Wahrscheinlichkeit nach nicht nur auf CG-Gehalt-abhängige lineare Weise erfolgt. Er kann jedoch keine gute Erklärung für das verblüffende Muster liefern, das durch Dehnen von GTG-gebänderten Chromosomen erhalten wird. Es ist vorstellbar, dass es durch Bestimmung der Abstände zwischen hellen und dunklen Chromosomenbänder in verschiedenen Kontraktionsstufen einmal möglich sein wird, einen Faktor oder eine Formel abzuleiten, deren Plausibilität in der Folge durch einen Vergleich der Bilder von natürlichen Chromosomen mit den entsprechenden virtuellen Chromosomen überprüft werden kann. Auch wenn es sein kann, dass die ungleichmäßige Verlängerung von kondensierten und dekondensierten dunklen und hellen Chromosomenbändern sogar visuell wahrnehmbar sind, besteht keine praktische Veranlassung für analytische Zwecke, da die präparationsabhängige Veränderung innerhalb der Chromosomenklasse selbst weiter ist als jene, die von der verzerrten künstlichen Dehnung mittels Computer resultiert.

35

Beispiel 4:

Virtuelle Chromosomen verknüpfen die zytogenetische und die molekulargenetische Karte

In Fig. 3a ist das ISCN-Chromosom 7 (Bandstufe 850) zusammen mit seinem virtuellen Gegenstück und drei verschiedenen Ideogrammen gezeigt. Das Bänderungsmuster des linken Ideogramms basiert auf der Lage der Wendepunkte zwischen CG-reicheren und CG-ärmeren Regionen im Sequenz-basierten virtuellen Chromosom. Die Kurve folgt dem mittleren CG-Gehalt. Das UCSC-Ideogramm (Ausgabe vom August 2001) ist in der Mitte und das entsprechende ISCN (Bandstufe 850) auf der rechten Seite platziert. Die linke Hälfte des virtuellen Chromosoms zeigt die rohen, unverstärkten Grauwerte des jeweiligen CG-Gehalts, wohingegen in der rechten Hälfte der Kontrast entsprechend der in Diagramm (b) dargestellten Kurve verstärkt ist. Die dünnen horizontalen Linien liefern eine absolute 10 Mb-Skala. Wie oben für Fig. 2 erläutert, könnte es jedoch sein, dass die DNA nicht auf solch lineare Weise entlang der Bänder verteilt ist, wie diese Skala nahe legt. Es wird auch klar, dass die Breite und Verteilung von Bändern in Ideogrammdarstellungen beträchtlich variieren kann, obwohl ihre Anzahl und Bezeichnung üblicherweise übereinstimmt. Es ist daher nicht möglich, jedes absolute oder relative chromosomale Auftreten auf irgendeinem Ideogramm genau zu positionieren. Virtuelle Chromosome lösen dieses Problem, indem sie die absolute Präzision der DNA-Sequenzpositionierung mit den willkürlichen Lageindikatoren irgendeiner Art Ideogramm verknüpfen, was durch die Linien verdeutlicht ist, die die Band-

55

grenzen der drei in dieser Figur gezeigten Beispiele kombinieren.

Beispiel 5:

Virtuelle in-situ-Hybridisierung

5 Da virtuelle Chromosome die DNA-Sequenz in sehr kondensierter Form symbolisieren, ist es nunmehr möglich, die genaue Position jeder Art von DNA-Sequenz oder Sequenzsatz unabhängig von der Anzahl und DNA-Sequenzlänge mit Hilfe der jeweiligen Nucleotid-Koordinaten mit einer bislang unbekannten zytogenetischen Präzision anzuzeigen. Ein Vergleichsbeispiel von konventioneller und virtueller FISH-Kartierung ist in Fig. 4 links für das MLL-Partnergen GRAF bei 5(q31) 10 gezeigt, dessen ursprüngliche, zytogenetisch bestimmte Lage zeichnerisch auf eine CGH-Software (Vysis, Doners, Grove, USA) beschränkt war. Das Metaphasenbild ist oben, das CGH-Kartierungsbild in der Mitte und das virtuelle Chromosom 5 mit dem verlängerten natürlichen unten gezeigt.

Bei früheren Versuchen, die zytogenetische Kartenlage in bestehende genomische Datenbanken zu integrieren, war man auf eine solche Anzeige von chromosomalen Ereignissen auf Ideogrammm-Koordinatensystemen angewiesen, da es nicht möglich war, diese beiden Datensätze direkt zu verknüpfen. Hier ist die Verteilung von 77 von 82 Chromosom-7-CCAP-BAC-Klonen als 15 Beispiel für die erzielbare Verbesserung bei der Zuordnung der absoluten und relativen Position eines ganzen Klonensatzes gezeigt.

Beispiel 6:

Konstruktion von Abnormitäten virtueller Chromosome

Die Definition eines speziellen chromosomalen Ereignisses mit molekularer Präzision erleichtert nunmehr die genaue Rekonstruktion jeder Chromosomen-Umlagerung mit einer bekannten molekularen Defektstellenlage. Wie hier durch das Beispiel der Translokation t (4;11) (q21;q23) 25 (s. Fig. 5) veranschaulicht, ist dies eine wichtige Voraussetzung für die potentielle Verwendung solcher virtueller Chromosomen-Umlagerungen in Mustererkennungssystemen und beim automatischen Karyotypisieren. Mit der aktuellen zytogenetischen Terminologie und Präzision kann die Lage der Defektstellen einer bestimmten Translokation nur willkürlich durch die Lage der betreffenden Bänder definiert werden. Je nach Sequenzfreisetzung umfasst das Band 4 (q21) zwischen 30 11,5 und 14,2 Mb (6,0 % - 7,4 % von Chromosom 4) und das Band 11 (q23) zwischen 10,9 und 11,7 Mb (7,9 % - 8,5 % von Chromosom 11). Ohne Kenntnis der genauen Position der beiden Gene AF4 und MLL, die infolge der Translokation zerstört und verschmelzt werden, innerhalb der Bänder könnten die Defektstellen irgendwo innerhalb dieser Bänder liegen. Zur Demonstration dieses Standpunkts wurden sie den äußeren Grenzen der fraglichen Bänder zugeordnet. Sehr 35 wahrscheinlich ist das auch bereits eine der höchsten Auflösungen, die mit einer durchschnittlichen morphologischen Chromosomenanalyse erzielt werden kann. Verglichen mit der Länge und den Bänderungsmustern von Ableitungschromosomen, die von exakt positionierten molekularen Defektstellen (durch * angedeutet) stammen, können jene, die aus undefinierten Defektstellenzuordnungen resultieren, ziemlich anders aussehen. Sie wären daher sicherlich unbrauchbar für Zwecke 40 des Vergleichs und neuerlichen Nachweises bei der Mustererkennung.

Beispiel 7:

Grafische Schnittstelle zwischen zytogenetischen und molekulargenetischen Datensätzen

Als Spitzeneinheiten der Humansequenz decken virtuelle Chromosome die neun Größenordnungen des kompletten Genoms auf stark kondensierte, leicht expandierbare und am natürlichsten 45 vorstellbare „morphologische“ Weise ab. Sie können daher als einzigartiges Vorfeld-Werkzeug für die Sichtbarmachung der in jeder Sequenz-Datenbank enthaltenen Informationen eingesetzt werden; nämlich im Prinzip von einem einzigen Basenpaar bis zu ganzen Chromosomen auf zytogenetische Weise. Als Beispiel sind hier in einer 1-Mb-Skala die Verteilung der annähernd 50 15.000 Gene und CpGs aus der UCSC-Datenbank entlang virtueller Chromosome und die entsprechenden UCSC-Farbideogramme gezeigt. Aus praktischen Gründen ist die Skala der Balken für die Gene nur in halber Höhe der CpG-Balken gezeigt und die Höhe dieser CpG-Balken am Chromosom 19 abgeschnitten.

Die vertikalen Balken auf der linken Seite der Chromosome zeigen die Größe und Lage von 55 heterochromatischen und Satellitenregionen an, die künstlich ergänzt wurden, da ihre Sequenz

derzeit noch nicht verfügbar ist. Die feinen horizontalen Balken auf der linken Seite der Chromosome deuten Sequenzlücken an.

5

PATENTANSPRÜCHE:

1. Verfahren zur Herstellung eines virtuellen Chromosoms, das ein entsprechendes natürliches Chromosom repräsentiert, **dadurch gekennzeichnet**, dass es die folgenden Schritte umfasst:
 - 10 - Unterteilen von Sequenzdaten des natürlichen Chromosoms in Fraktionen mit einer Länge von mindestens 10.000 bp,
 - Bestimmen des CG-Gehalts in jeder Fraktion,
 - Berechnen eines Werts zwischen einem Minimalwert und einem Maximalwert für jede Fraktion gemäß dem CG-Gehalt, und
 - 15 - Herstellen des virtuellen Chromosoms durch Darstellen jeder Fraktion mit dem Wert.
2. Verfahren nach Anspruch 1, **dadurch gekennzeichnet**, dass der Wert ein Lichtwert ist.
3. Verfahren nach Anspruch 2, **dadurch gekennzeichnet**, dass der Maximalwert weiß, der Minimalwert schwarz und Werte dazwischen in Grauschattierungen dargestellt werden.
- 20 4. Verfahren nach einem der Ansprüche 1 bis 3, **dadurch gekennzeichnet**, dass das natürliche Chromosom in Fraktionen einer Länge von 10.000 bis 1.000.000 bp, vorzugsweise einer Länge von 50.000 bis 500.000 bp, noch bevorzugter einer Länge von 100.000 bis 300.000 bp unterteilt wird.
5. Verfahren nach einem der Ansprüche 1 bis 4, **dadurch gekennzeichnet**, dass die Fraktion mit einem CG-Gehalt von 30 bis 35 %, vorzugsweise 33 %, einem Minimalwert und die
 - 25 Fraktion mit einem CG-Gehalt von 60 bis 65 %, vorzugsweise 62 %, einem Maximalwert zugeordnet wird.
6. Verfahren nach einem der Ansprüche 1 bis 5, **dadurch gekennzeichnet**, dass Fraktionen mit unbekannter Sequenz der Wert gemäß ihrer morphologischen Erscheinung zugeordnet wird.
- 30 7. Verfahren nach einem der Ansprüche 1 bis 6, **dadurch gekennzeichnet**, dass nach der Herstellung des virtuellen Chromosoms ein Filter zur Glättung der Erscheinung, vorzugsweise ein Gaußsches Faltungsfiter, angewendet wird.
8. Verfahren nach einem der Ansprüche 1 bis 7, **dadurch gekennzeichnet**, dass zur Herstellung des virtuellen Chromosoms ein Skalenkorrekturfilter angewendet wird.
- 35 9. Virtuelles Chromosom oder Teil davon, das bzw. der durch Werte gemäß seinem CG-Gehalt dargestellt wird, **dadurch gekennzeichnet**, dass die Herstellung nach einem der Ansprüche 1 bis 6 erfolgt.
10. Virtuelles Chromosom oder Teil davon nach Anspruch 9, **dadurch gekennzeichnet**, dass der Wert ein Lichtwert ist.
- 40 11. Virtuelles Chromosom oder Teil davon nach Anspruch 10, **dadurch gekennzeichnet**, dass der Maximalwert weiß, der Minimalwert schwarz und Werte dazwischen in Grauschattierungen dargestellt sind.
12. Satz von virtuellen Chromosomen oder Teilen davon, **dadurch gekennzeichnet**, dass er
 - 45 zwei oder mehr Chromosome oder Teile davon nach einem der Ansprüche 9 bis 11 umfasst.
13. Satz nach Anspruch 12, **dadurch gekennzeichnet**, dass er Chromosome oder Teile davon umfasst, die für einen oder mehrere Organismen spezifisch sind.
14. Satz nach Anspruch 13, **dadurch gekennzeichnet**, dass er 24 menschliche Chromosome oder Teile davon umfasst.
- 50 15. Satz nach Anspruch 14, **dadurch gekennzeichnet**, dass er weiters zusätzliche modifizierte Chromosome oder Teile davon, vorzugsweise Chromosome mit Translokationen, umfasst.
16. Verwendung eines Satzes von virtuellen Chromosomen nach einem der Ansprüche 12 bis 15 zur Katalogisierung von Chromosomenmodifikationen.
- 55 17. Verwendung eines Satzes von virtuellen Chromosomen nach einem der Ansprüche 12 bis

- 15 zur virtuellen Kartierung der chromosomalen Position einer Sequenz.
18. Verwendung eines Satzes von virtuellen Chromosomen nach einem der Ansprüche 12 bis 15 als Schnittstelle zwischen morphologischen und molekulargenetischen Daten.
 - 5 19. Verwendung nach Anspruch 18, **dadurch gekennzeichnet**, dass die morphologischen Daten von Informationen stammen, die auf dem Internationalen System für Humane Zytogenetische Nomenklatur (ISCN) basieren.
 20. Verwendung nach einem der Ansprüche 16 bis 19, **dadurch gekennzeichnet**, dass der Chromosomensatz als Referenz zur Klassifizierung eines Phänotyps zu einer Sequenzanordnung dient.
 - 10 21. Verwendung nach einem der Ansprüche 16 bis 20, **dadurch gekennzeichnet**, dass der Chromosomensatz als Werkzeug zur Durchführung von Struktur- bzw. Funktionsanalysen einer Sequenzanordnung dient.
 - 15 22. Verwendung nach einem der Ansprüche 16 bis 21, **dadurch gekennzeichnet**, dass der Chromosomensatz als Werkzeug zur Bestimmung des Einflusses eines bestimmten Faktors auf eine Sequenzanordnung dient.

HIEZU 6 BLATT ZEICHNUNGEN

20

25

30

35

40

45

50

55

FIG.1

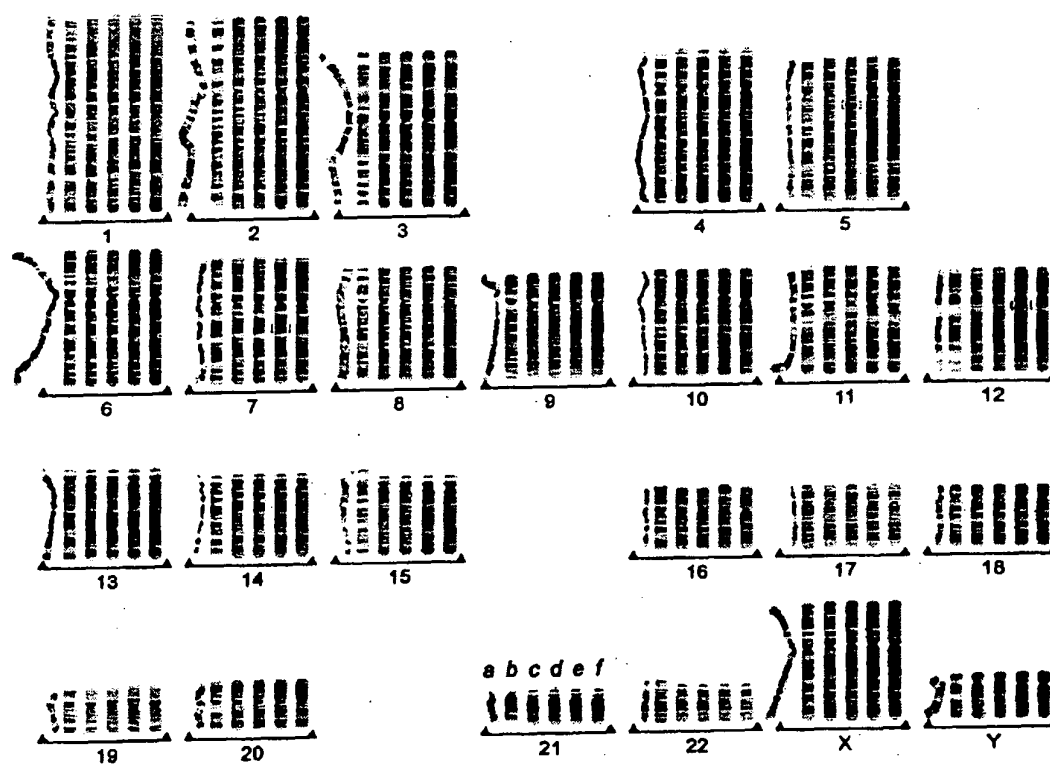


FIG. 2

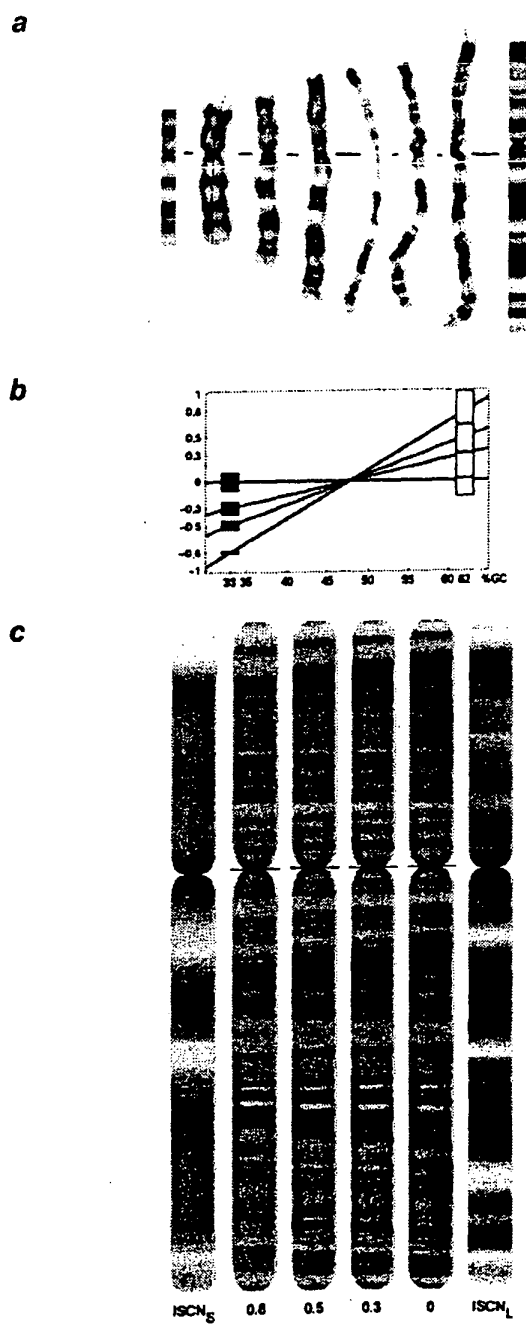


FIG. 3

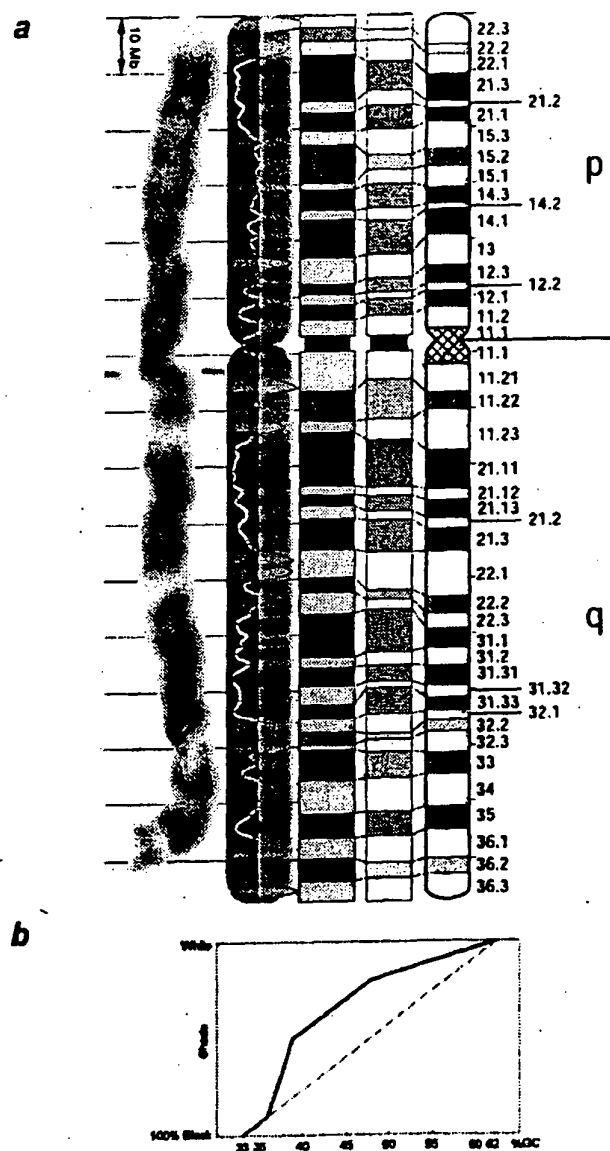
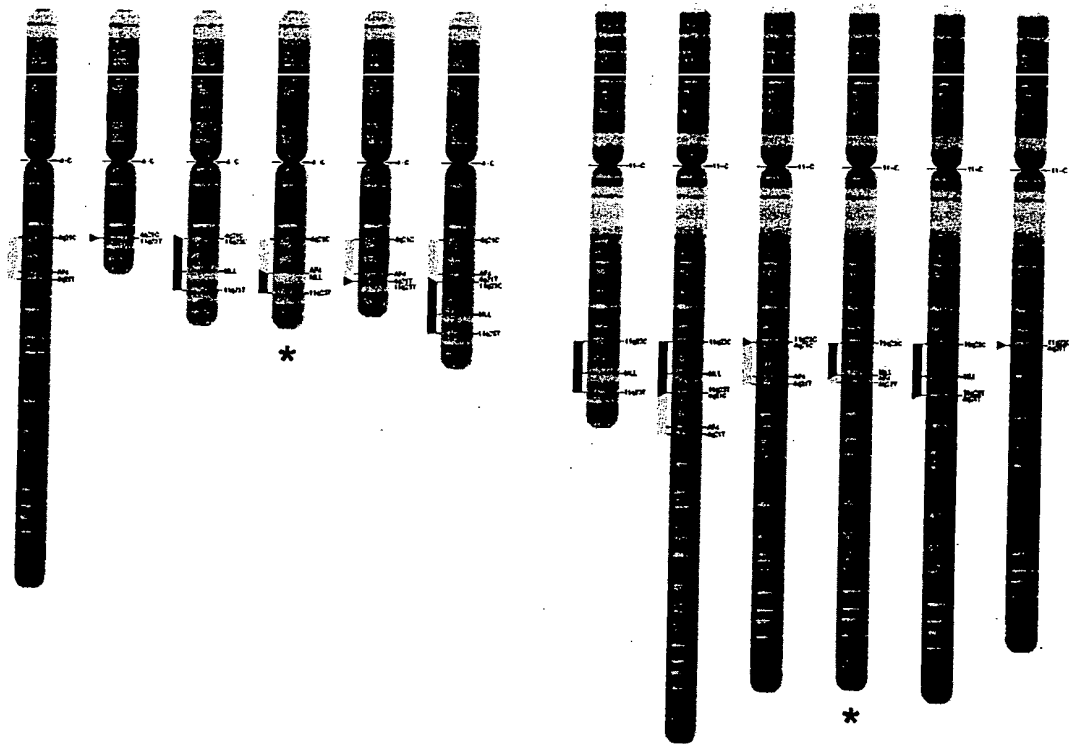


FIG. 4



FIG. 5



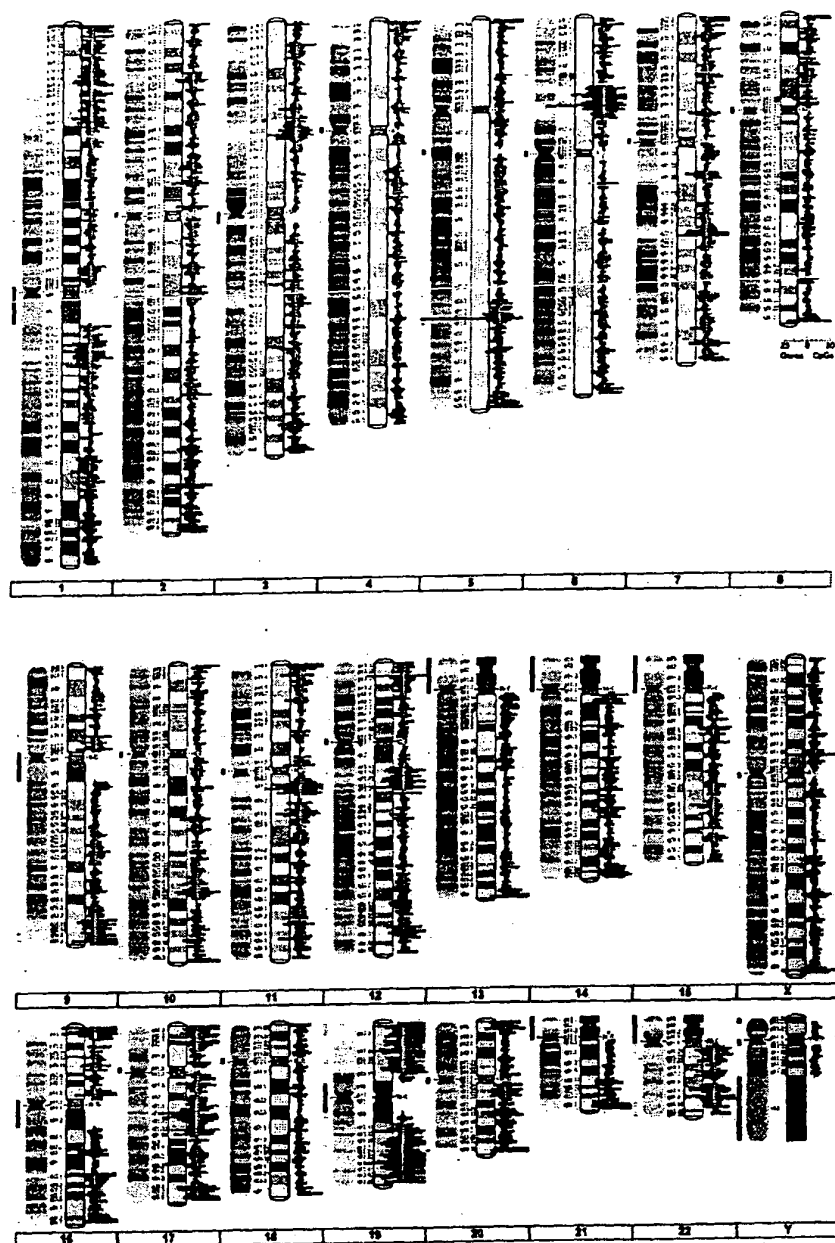


FIG. 6